中国地质灾害与防治学报
THE CHINESE JOURNAL OF GEOLOGICAL HAZARD AND CONTROL
中国科技核心期刊
《中国科学引用文数据库》来源期刊
Caj-cd规范获奖期刊

基于SMOTE-Tomek和CNN耦合的滑坡易发性评价模型及其应用

于宪煜，汤　礼

Landslide susceptibility mapping model based on a coupled model of SMOTE-Tomek and CNN and its application: A case study in the Zigui-Badong section of the Three Gorges Reservoir area

YU Xianyu and TANG Li

在线阅读 View online: https://doi.org/10.16031/j.cnki.issn.1003-8035.202212002

DOI: 10.16031/j.cnki.issn.1003-8035.202212002
于宪煜, 汤礼. 基于 SMOTE-Tomek 和 CNN 耦合的滑坡易发性评价模型及其应用——以三峡库区秭归—巴东段为例[J]. 中国地质灾害与防治学报, 2024, 35(3): 141-151.
YU Xianyu, TANG Li. Landslide susceptibility mapping model based on a coupled model of SMOTE-Tomek and CNN and its application: A case study in the Zigui-Badong section of the Three Gorges Reservoir area[J]. The Chinese Journal of Geological Hazard and Control, 2024, 35(3): 141-151.

基于 SMOTE-Tomek 和 CNN 耦合的滑坡易发性评价模型及其应用

——以三峡库区秭归—巴东段为例

于宪煜, 汤　礼
（湖北工业大学土木建筑与环境学院，湖北 武汉　430068）

摘要: 中国是受滑坡灾害影响较为严重的国家, 滑坡对受灾害影响地区的人民生命与财产造成了巨大的威胁。滑坡易发性评价作为对滑坡风险预测的重要工具, 具有重要的防灾减灾的意义, 但是传统的滑坡易发性评价中存在滑坡与非滑坡样本数据不平衡的问题, 使得训练集的建立在本质上是对非滑坡数据进行了欠采样, 导致滑坡事件的重要信息特征丢失, 进而影响到滑坡易发性评价的可靠性。文章以三峡库区巴东至秭归段为例, 选取高程、坡度等 14 个评价因子作为滑坡易发性评价因子, 划分原始训练集与验证集, 采用 SMOTE-Tomek 方法 (synthetic minority oversampling technique-Tomek Links, SMOTE-Tomek) 处理原始训练数据集, 构建输入训练集, 输入并训练卷积神经网络模型 (convolutional neural networks, CNN), 得到 SMOTE-Tomek-CNN 耦合模型, 再通过将 SMOTE-Tomek 方法与传统的欠采样方法 (random undersampling, RUS), 分别与 CNN 模型和支持向量机模型 (support vector machine, SVM) 交叉组合成 SMOTE-Tomek-SVM、RUS-CNN 和 RUS-SVM 三种耦合模型, 并与 SMOTE-CNN 耦合模型进行对比。结果表明, 在四种耦合模型中, SMOTE-CNN 耦合模型的特定类别精度与 ROC 曲线下面积较高, 结果分别为 73.60% 和 0.965, 表明该方法的预测能力优于传统的方法, 能为研究区滑坡预测工作提供可靠参考。

关键词: 滑坡; 滑坡易发性评价; SMOTE-Tomek; 卷积神经网络; 不平衡数据
中图分类号: P642.22　　**文献标志码:** A　　**文章编号:** 1003-8035(2024)03-0141-11

Landslide susceptibility mapping model based on a coupled model of SMOTE-Tomek and CNN and its application: A case study in the Zigui-Badong section of the Three Gorges Reservoir area

YU Xianyu，TANG Li
（*School of Civil Engineering Architecture and Environment*, *Hubei University of Technology*, *Wuhan*, *Hubei*　430068, *China*）

Abstract: China is a nation severely impacted by landslide disasters, which poses a great threat to the lives and properties of people in the disaster-affected areas. Landslide susceptibility assessment, as an important tool for landslide risk prediction, is of great significance for disaster mitigation and prevention. However, traditional landslide susceptibility assessment faces the issue

收稿日期: 2022-12-03; **修订日期:** 2023-03-17　　**投稿网址:** https://www.zgdzzhyfzxb.com/
基金项目: 国家自然科学基金青年项目(41807297)
第一作者: 于宪煜(1987—), 男, 湖北武汉人, 博士, 副教授, 主要研究方向为滑坡地质灾害分析与预测。E-mail: yuxianyu@hbut.edu.cn

of imbalanced data between landslide and non-landslide samples, leading to the inherent undersampling of non-landslide data in the training set. This results in the loss of important information features related to landslide events, thereby affecting the reliability of landslide susceptibility assessment. In this study, using the Zigui-Badong section of the Three Gorges Reservoir Area as an example, 14 evaluation factors, such as elevation and slope were chosen as landslide susceptibility assessment factors, and the original training set and the validation set were divided. In this study, the synthetic minority oversampling technique - Tomek Links (SMOTE-Tomek) method was employed to process the original training dataset, construct the input training set. A convolutional neural networks (CNN) was then trained using this input data, resulting in the SMOTE-Tomek-CNN coupling model. In addition, by intersecting the SMOTE-Tomek method with undersampling methods (random undersampling, RUS), they were separately coupled with the CNN model and support vector machine model (SVM) to form three coupled models: SMOTE-Tomek-SVM, RUS-CNN, and RUS-SVM. These were compared with the SMOTE-CNN coupled model. The results indicate that, among the four coupling models, the SMOTE-CNN coupled model has higher specific class accuracy and area under the ROC curve, with values of 73.60% and 0.965, respectively. This indicates that this method's predictive ability is superior to that of traditional methods, making it a reliable resource for landslide prediction in the studied area.

Keywords: landslide；landslide susceptibility assessment；SMOTE-Tomek；convolutional neural network；unbalanced data

0 引言

滑坡易发性评价是以工程地质类比法为理论基础，可以对研究区域内的滑坡空间分布进行预测的一种方法[1]。这种方法可以根据理论基础的不同，分为确定性方法与非确定方法。当前滑坡易发性评价常用的非确定性方法有支持向量机（support vector machine, SVM）[2－3]，Logistic 回归[4－6]，决策树[7]与神经网络[8]等。

深度学习属于机器学习研究中的一个新领域，通过构建深层的网络结构，拥有很强的非线性拟合能力。已有学者将卷积神经网络（convolutional neural networks, CNN）用于滑坡易发性评价分析之中，以现有的滑坡数据挖掘了主要致灾因子和成灾规律，并验证了这种方法的可行性[9]。

上述机器学习方法在滑坡易发性评价中均得到了成功地应用，但是需要注意的是，这些方法中训练模型的数据集是基于滑坡数据与非滑坡数量是均衡的假定，但在实际情况中数据不平衡问题是普遍存在的，即研究区内的非滑坡的面积是远远大于滑坡的面积[10]。传统的训练数据集构建过程使得模型在训练时被动地丢失掉非滑坡数据内的重要信息特征，从而进一步影响了滑坡易发性评价的可靠性。

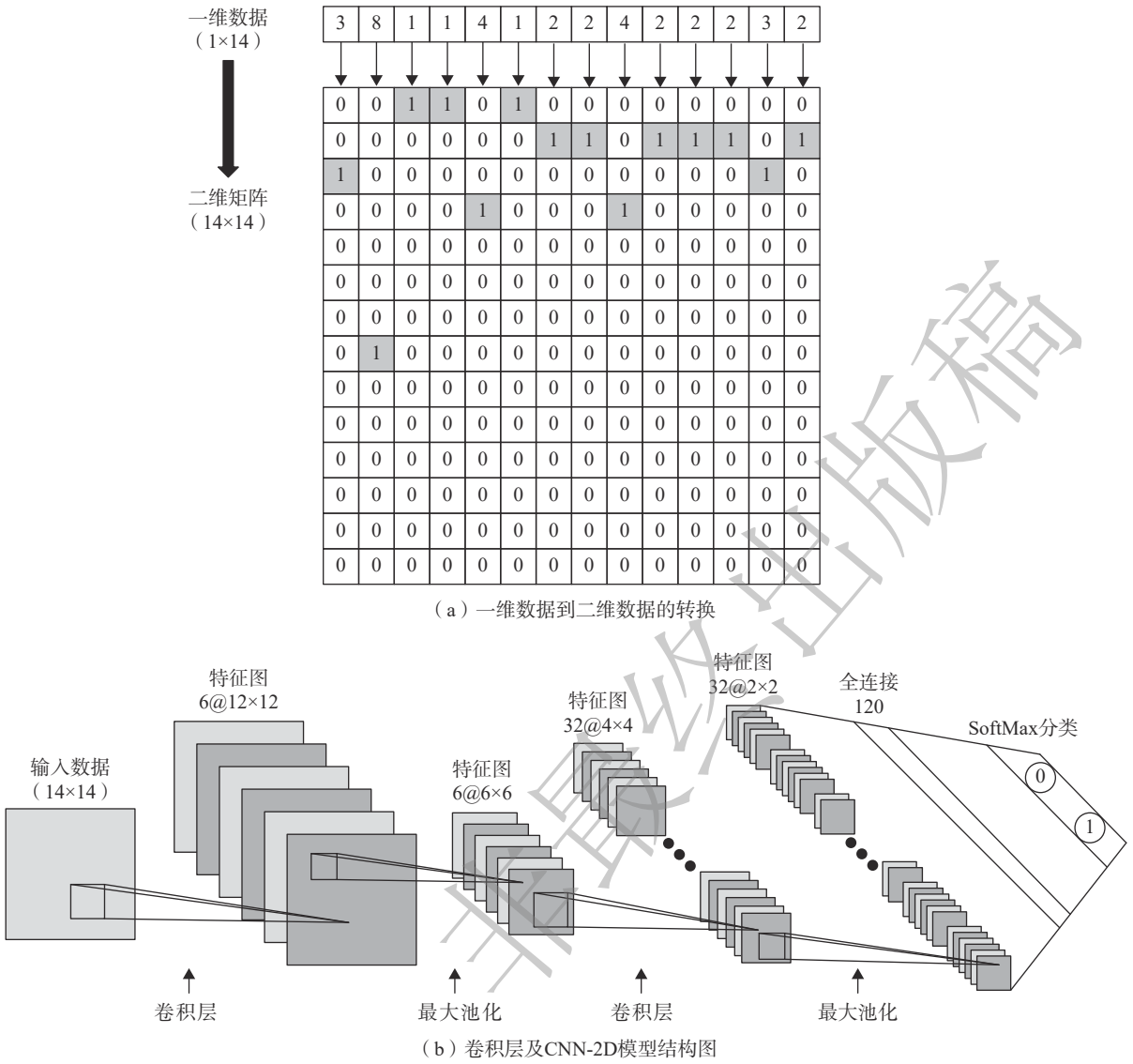
本文综合前人对三峡库区秭归到巴东段滑坡易发性评价的研究基础，以及地质学、地貌学、统计分析和机器学习等多学科的理论方法，SMOTE-Tomek 方法（synthetic minority oversampling technique-Tomek Links, SMOTE-Tomek）与 CNN 模型耦合应用于滑坡易发性评价，同时引入 SVM 模型作为对照，客观比较与评价该采样方法与不同机器学习方法组合所得到的结果，使得滑坡数据不平衡问题对滑坡易发性评价的影响最小化，提高其预测结果的精确性和可靠性，帮助相关部门顺利开展防灾减灾的工作，以减少滑坡灾害给生产生活带来的损失。

1 研究方法

1.1 卷积神经网络

Lecun 利用梯度下降更新参数的思想设计了 CNN，CNN 作为一种强大的深度学习技术，在不需要对输入数据进行分类操作的前提下，能够自主学习海量输入数据与输出数据之间的潜在规则，提取数据的局部特征，从而进行高精度分类。本研究使用一种 CNN-2D 结构，实现此结构，需将输入的一维数据转化为二维矩阵，具体来说，一维数据中的每一个因子与二维矩阵中每一个列向量对应，对于这个列向量，对应于相应属性值的位置处的值被赋值为 1，其他值被赋值为 0。矩阵大小为滑坡易发性评价因子的个数，本研究选取 L 个因子，故该二维矩阵的大小为 $L\times L$，通过独热编码以及补零等方法完成了到二维矩阵的转换操作，该过程如图 1(a)所示。

CNN-2D 结构有两个内核大小都为 $m\times m$ 卷积层，和两个内核大小为 $n\times n$ 的最大池化层。假设将输入的滑坡栅格单元数据转化为一个 $a\times a$ 的二维矩阵，输入第一个卷积层中，得到 N 个 $(a-m+1)\times(a-m+1)$ 的特征图。最大池化层紧接在卷积层后使用，大小为 $n\times n$，该层输



（a）一维数据到二维数据的转换

（b）卷积层及CNN-2D模型结构图

图 1　一维数据到二维矩阵的转换及 CNN-2D 模型结构图

Fig. 1　Transformation from one-dimensional data to a two-dimensional matrix and structure diagram of the CNN-2D model

出大小为$[(a-m+1)/n]\times[(a-m+1)/n]$的特征图，此后的卷积层和最大池化层重复上述过程，最终输出 M 个特征图，大小为$[(a-(n+1)(m-1)/n^2]\times[(a-(n+1)(m-1)/n^2]$。最后一个最大池化层后有一个与所有神经单元完全连接的全连接层，将特征图展开为向量并重新组织提取的特征。最后在输出层上的两个神经单元用 1 和 0 分别表示滑坡与非滑坡，该结构比传统机器学习模型有更高的精度[11]。该结构如图 1(b)所示，在此结构中，a = 14，m = 3，n = 2，N = 6，M = 32。

1.2　支持向量机模型

支持向量机的原理是构建一个 n-维超平面作为分类平面，对输入的数据进行分类。假设一个非线性可分的向量 $x_i(i=1,2,\cdots,n)$，包含了两类 $y_i=\pm1$，则这个 n-维超平面定义式如式(1)所示：

$$\begin{cases}\min\dfrac{1}{2}\|w\|^2\\ s.t.y_i((w\times x_i)+b)\geqslant 1\end{cases}\tag{1}$$

式中：$\|w\|$——w 的 2-范数；

w——垂直与超平面的向量；

x_i——超平面上的点；

b——为了使得超平面不会通过坐标轴原点的常数。

训练数据集通过核函数 $K(x_i,x_j)$ 转换到 n-维空间中，这个核函数本质上是一个映射函数。文献[12]表明使用径向基核 RBF 核的支持向量机模型的性能是优于其他核支持向量机模型，故本研究采用基于 RBF 核的支持向量机进行滑坡易发性评价。

1.3 SMOTE-Tomek 方法

SMOTE 是 Chawla 等[13]提出的一种过采样算法，通过对原始数据进行函数运算生成少数类的数据，其过程为：(1)对于一个少数类的数据 x_i，计算它到其他少数类数据集中所有数据的距离，得到其 k 个近邻；(2)对于每一个少数类数据 x_i，从其 k 近邻中随机选择若干个数据，假设选择的近邻为 $\tilde{x}$；(3)对于每一个随机选出的近邻 $\tilde{x}$，分别与原数据按照式(2)构建新的数据。

$$x_{\text{new}} = x + \text{rand}(0,1) \times (\tilde{x} - x) \tag{2}$$

式中：x_{new}——新构建的少数类数据；

rand(0, 1)——0 到 1 之间的一个随机数，该随机数是在区间内离散均匀分布的伪随机数，所以每个新生成的数据所使用的随机数是不同的，见图 2(a)(b)。

滑坡样本与非滑坡样本的不平衡会影响模型的准确性，常通过使用欠采样方法在非滑坡样本中抽取部分样本与滑坡样本数量达到均衡，或者使用过采样方法增加滑坡样本与非滑坡样本数量达到均衡[14]。在过采样方法中，SMOTE 算法容易产生数据重叠问题，而 Tomek Links 方法可以在一定程度上缓解此问题[15]，本研究将二者组合成为的 SMOTE-Tomek 方法对不平衡的滑坡数据进行处理。Tomek Links 方法是计算所属不同类别的两个数据实例 x_i 和 x_j 两者之间的距离，两者的距离用 $d(x_i, x_j)$ 表示。如果数据集中不存在除 x_i 和 x_j 之外的一个其他数据点 x，满足 $d(x_i, x)<d(x_i, x_j)$ 或者 $d(x_j, x)<d(x_i, x_j)$ 的条件时，则 x_i 和 x_j 被称作为 Tomek Links 对。如果两个点被判断为 Tomek Links 对，则说明这两者中含有一个是噪声数据，或者表明两者都是在边界位置上。在经过 SMOTE 方法处理过的数据集使用 Tomek Links 方法删除掉 Tomek Links 对，见图 2(c)(d)。

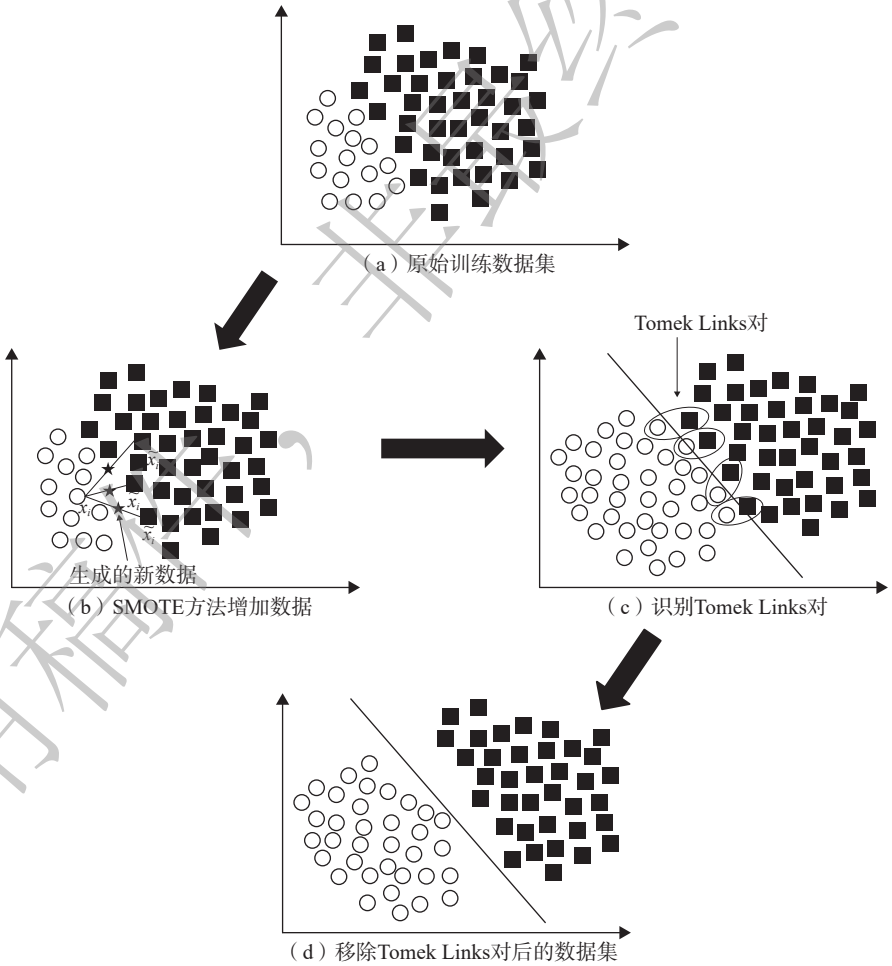


图 2 SMOTE-Tomek 方法处理数据集

Fig. 2 Processing of the dataset using the SMOTE-Tomek method

2　研究区概况及数据源

研究区位于三峡库区秭归到巴东段，该段位于重庆奉节以东，宜昌秭归以西，东西跨越约 54 km，南北跨越约 16 km(图 3)。由于在长江长年下切作用，导致研究区内整个地形地貌有明显的四周高和中间低的盆地特征，沿江两岸的地势表现出中间低，两岸高[16]。该区域处于中纬度区域，属于亚热带季风气候，气候和降雨量随季节变化明显，同时气温受高差影响变化明显，巴东县年平均降雨量为 1 034.3 mm，秭归地区年平均降雨量为 1 158.9 mm。

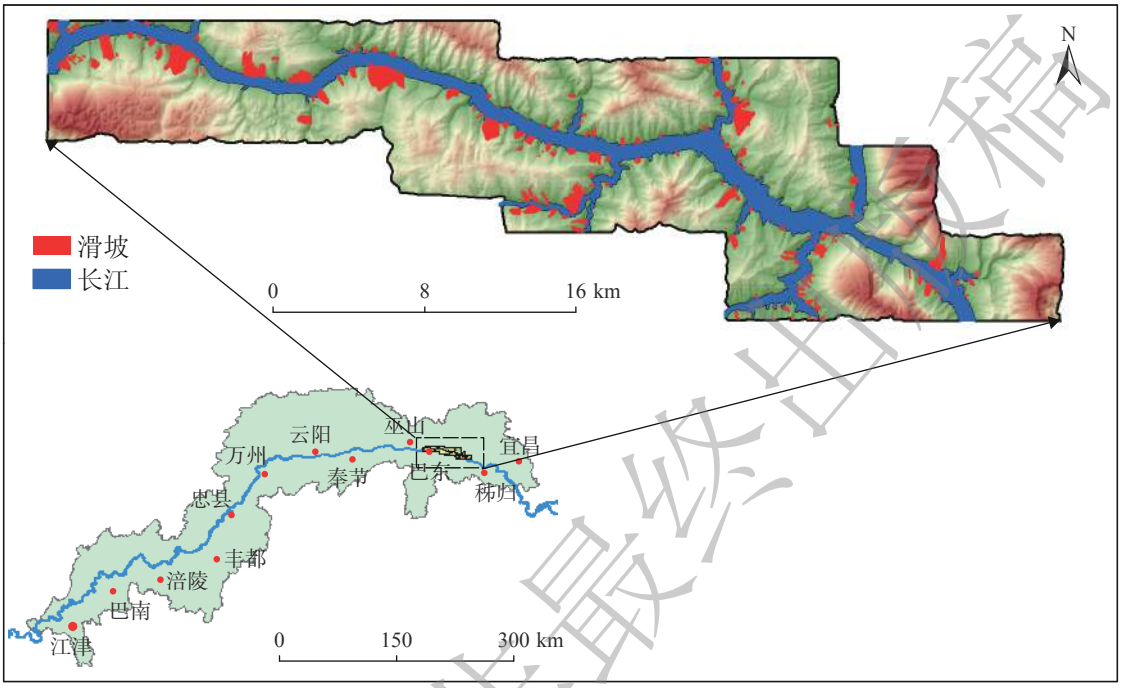


图 3　研究区地理位置及滑坡分布情况

Fig. 3　Geographic location and landslide distribution in the study area

有学者[17]将滑坡易发性评价的计算单元总结为归纳为 5 种，它们分别是：子流域单元、斜坡单元、唯一条件单元、地域单元以及栅格单元。其中栅格单元的优势是可利用栅格数据本身的像元作为计算单元，这样能保证每个计算单元的面积是相同的，并且它是适合输入 CNN-2D 结构的所需的数据形式，栅格故本研究采用 30 m×30 m 的栅格单元作为评价单元。数据源主要有：(1)1∶10 000 比例尺的滑坡灾害图；(2)1∶50 000 比例尺的地形图和 1∶50 000 比例尺的地质图；(3)Landsat-8 卫星 OLI 传感器数据；(4)中国气象局大气降水数据；(5)高级星载热发射和发射辐射计全球数字高程模型数据。

根据调查结果显示新生滑坡和复活的古滑坡共计 202 处滑坡，分布如图 3 所示，总面积为 23.4 km^2，占研究区域面积的 6.03%。

3　滑坡易发性评价因子的筛选

通过分析前人的研究成果[18]，结合数据源选取高程、坡向、坡度、坡长、地形表面纹理、地形起伏度指数、距断层距离、岩性、距长江距离、地形湿度指数、年平均降雨量、土地利用类型、归一化植被指数和距道路距离共 14 个因子作为滑坡易发性评价因子，其分级情况如表 1 所示。其中大多具有天然相关性，进而要对这 14 个因子进行分析和筛选，步骤如下：

表 1　14 个因子多重共线性分析

Table 1　Multicollinearity analysis of 14 factors

因子	*TOL*	*VIF*	因子	*TOL*	*VIF*
高程	0.363	2.758	岩性	0.776	1.289
坡向	0.971	1.030	距长江距离	0.388	2.578
坡度	0.118	8.454	地形湿度指数	0.838	1.193
坡长	0.631	1.586	年平均降雨量	0.511	1.957
地形表面纹理	0.887	1.274	土地利用类型	0.862	1.160
地形起伏度指数	0.114	8.734	归一化植被指数	0.750	1.334
距断层距离	0.818	1.223	距道路距离	0.513	1.951

(1)使用皮尔逊相关系数分析(Pearson correlation coefficient, PCC)。为去除因子间相关性对模型预测的

影响，利用皮尔逊相关系数分析对 14 个因子组合进行分析，分析结果如图 4 所示，所有因子组合的分析结果均通过相关性检测。

（2）使用容忍度（*TOL*）及方差膨胀因子（*VIF*）进行多重共线性分析。多重共线性分析结果如表 2 所示，最大的方差膨胀因子值是 8.734，满足 *VIF*<10，且 *TOL*>0.1，故本研究选取的滑坡易发性评价因子之间不存在多重共线性。

（3）利用 Relief-F 算法的因子重要性筛选。Relief-F 方法可以计算滑坡易发性评价因子与滑坡之间的相关性来评估该因子，以确定该因子对滑坡发生的相对重要性[19]。Relief-F 随机选择一个数据 R，并且使用数据标签为 R 的 k 最近邻数据和来自 R 的不同标签分别构建数据集 H 和 M，对于所有特征 w_i，按照式（3）更新特征的权重：

$$w_i = w_i - \sum_{j=1}^{k} \frac{diff(A_i, R, H_i)}{mk} + \sum_{C \neq Class(R)} \left\{ \frac{p(C)}{1 - p[Class(R)]} \sum_{j=1}^{k} \frac{diff[A_i, R, M_j(C)]}{mk} \right\} \tag{3}$$

式中：C——数据标签；

$p(C)$——C 类的概率；

R 类——R 类的数据标签，是 C 类的第 j 个数据；

$diff(A_i,R,H_i)$ 和 $diff[A_i,R,M_j(C)]$——距离函数，将在重复该计算过程 m 次后计算因子的重要性。

各因子的 Relief-F 系数如图 5 所示。

Relief-F 系数最低的是归一化植被指数因子，其

表 2　选取的滑坡易发性评价因子

Table 2　The selected factors for the landslide susceptibility assessment

因子	分级	因子	分级	因子	分级
高程/m	<400	地势起伏度指数	0 ~ 35	土地利用类型	水体
	400 ~ 800		35 ~ 70		森林
	800 ~ 1 200		70 ~ 105		人工覆盖面
	1 200 ~ 1 600		105 ~ 140		草地
	>1 600		>140		农业用地
坡向	平地	距断层距离/m	0 ~ 1 500	归一化植被指数	<0.075
	正北		1 500 ~ 3 000		0.075 ~ 0.15
	北东		3 000 ~ 4 500		0.15 ~ 0.225
	正东		4 500 ~ 6 000		0.225 ~ 0.3
	南东		6 000 ~ 7 500		0.3 ~ 0.375
	正南		>7 500		>0.375
	西南	岩性	硬岩	距道路距离/m	0 ~ 800
	正西		软岩		800 ~ 1 600
坡度/(°)	0 ~ 15		软硬交替		1 600 ~ 2 400
	15 ~ 30	距长江距离/m	0 ~ 1 000		2 400 ~ 3 200
	30 ~ 45		1 000 ~ 2000		3 200 ~ 4 000
	45 ~ 60		2 000 ~ 3 000		>4 000
	60 ~ 75		3 000 ~ 4 000		
	>75		4 000-5 000		
坡长/m	0 ~ 800		>5 000		
	800 ~ 1 600	地形湿度指数	<6		
	1 600 ~ 2 400		6 ~ 9		
	2 400 ~ 3 200		9 ~ 12		
	>3 200		12 ~ 15		
地形表面纹理	0 ~ 0.14		15 ~ 18		
	0.14 ~ 0.28	年平均降雨量/mm	<990		
	0.28 ~ 0.42		990 ~ 1 020		
	0.42 ~ 0.56		1 020 ~ 1 050		
	>0.56		1 050 ~ 1 080		
			1 080 ~ 1 110		
			>1 110		

	高程	坡向	坡度	坡长	地形表面纹理	地形起伏度指数	距断层距离	岩性	距长江距离	地形湿度指数	年平均降雨量	土地利用类型	归一化植被指数	距道路距离
高程	1	0.42	0.074	−0.136	0.159	0.085	−0.125	−0.275	0.512	0.135	−0.174	−0.154	0.338	0.598
坡向	0.42	1	0.011	−0.011	0.03	0.008	−0.036	0.008	0.011	0.142	0.028	−0.043	−0.005	0.014
坡度	0.074	0.011	1	0.013	−0.171	0.0939	0.058	−0.073	−0.065	0.113	−0.436	−0.206	0.03	0.11
坡长	−0.136	−0.011	0.013	1	−0.28	0.000586	−0.012	−0.053	−0.106	0.049	0.524	0.034	−0.035	−0.051
地形表面纹理	0.159	0.03	−0.171	−0.28	1	−0.107	−0.056	0.099	0.274	−0.072	−0.145	0.000964	0.143	0.11
地形起伏度指数	0.085	0.008	0.0939	0.000586	−0.107	1	0.029	−0.088	−0.046	0.125	−0.409	−0.202	0.033	0.119
距断层距离	−0.125	−0.036	0.058	−0.012	−0.056	0.029	1	0.291	−0.084	−0.233	−0.042	0.067	0.046	0.107
岩性	−0.275	0.008	−0.073	−0.053	0.099	−0.088	0.291	1	−0.149	−0.166	0.006	0.131	0.076	−0.009
距长江距离	0.512	0.011	−0.065	−0.106	0.274	−0.046	−0.084	−0.149	1	0.04	−0.06	0.009	0.389	0.582
地形湿度指数	0.135	0.142	0.113	0.049	−0.072	0.125	−0.233	−0.166	0.04	1	0.007	−0.137	−0.125	0.162
年平均降雨量	−0.174	0.028	−0.436	0.524	−0.145	−0.409	−0.042	0.006	-0.06	0.007	1	0.123	−0.018	−0.091
土地利用类型	−0.154	−0.043	−0.206	0.034	0.000964	−0.202	0.067	0.131	0.009	−0.137	0.123	1	0.176	−0.104
归一化植被指数	0.338	−0.005	0.03	−0.035	0.143	0.033	0.046	0.076	0.389	−0.125	−0.018	0.176	1	0.221
距道路距离	0.598	0.014	0.11	−0.051	0.11	0.119	0.107	−0.009	0.582	0.162	−0.091	−0.104	0.221	1

图 4　14 个因子的 PCC 系数矩阵

Fig. 4　Pearson correlation coefficient (PCC) matrix for the 14 factors

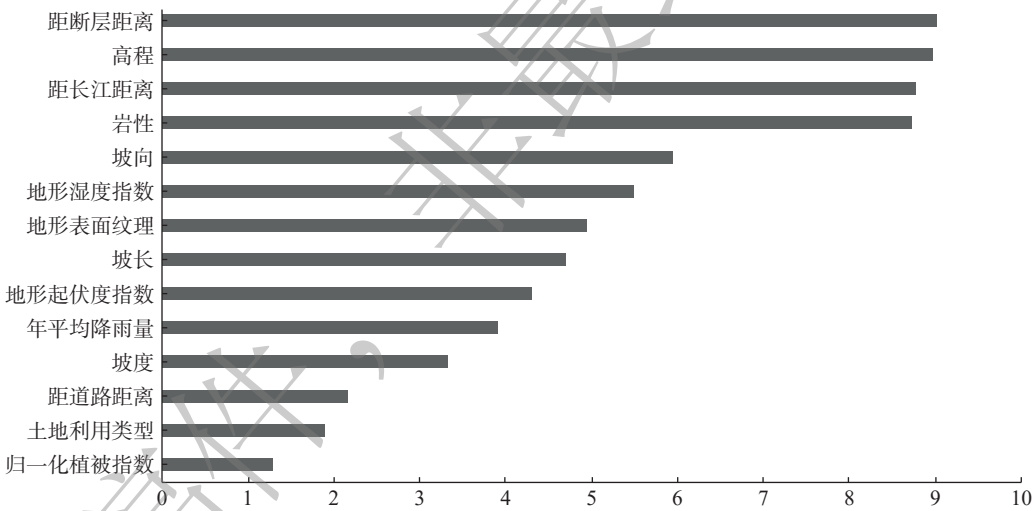


图 5　14 个因子的 Relief-F 系数

Fig. 5　Relief-F coefficients for the 14 factors

值大于 0，则表示所选取的因子对于滑坡发生都是重要的。

4　滑坡易发性评价结果与分析

4.1　训练数据集的建立

据统计研究区共计有 425 257 个栅格单元。随机选取 70% 的滑坡数据，即 141 处滑坡（19 263 个栅格单元）和 70% 的非滑坡数据（279 736 个栅格单元）构建原始训练数据集。先使用 SMOTE 方法增加原始训练数据集中的滑坡数据，生成与非滑坡数据相同数量的滑坡栅格单元，数据集包含 559 472 个栅格单元，接着使用 Tomek Links 方法在数据集中找到 112 个 Tomek Links 对（即 224 个栅格单元），并将其在数据集中删除掉，最终训练数据集的栅格单元数量为 559 248。

4.2　四种耦合模型的建立

将上一节得到的训练数据集导入 CNN 模型进行训练与建模，组成 SMOTE-Tomek-CNN 耦合模型。同时使用传统欠采样（random undersampling, RUS）处理得到

的一组平衡数据集，将其与 SMOTE-Tomek 方法得到的训练数据集分别与 CNN 模型与 SVM 模型交叉耦合，即得到 RUS-CNN、RUS-SVM 和 SMOTE-Tomek-SVM 三种耦合模型，建立对比试验。

CNN 模型各项参数设置如表 3 所示。其中，卷积层选取的激活函数采用的 ReLu 函数，ReLu 函数能在一定程度上加快模型的收敛速度，并可以在一定程度上克服梯度消失的问题[20]。交叉熵误差可以真实地反映出分类结果和预测结果的误差，常和 Softmax 分类一起使用将回归变成概率分布。CNN 模型在训练过程中，需要迭代更新参数以提高模型分类效果，本研究使用的权重更新算法为 Adam 算法，它是随机梯度下降算法的扩展，它能有效减少计算机资源的消耗和降低对参数的调整要求[21]。

表 3 CNN 模型参数设置表

Table 3 Configuration of parameters for the CNN model

CNN-2D各参数项	参数值	CNN-2D各参数项	参数值
卷积核大小	3 × 3	优化器	Adam
最大池化核	2 × 2	迭代次数	20
激活函数	ReLu	批量数据大小	2 000
误差函数	交叉熵误差	学习率	0.001

4.3 四种耦合模型的滑坡易发性评价结果

将全部数据导入训练好的 CNN 模型与 SVM 模型得到研究区内每个栅格单元的滑坡易发性指数。

为提高滑坡易发性指数的可读性，以及完整了解滑坡易发性的分布，根据 0 ~ 0.5、0.5 ~ 0.75、0.75 ~ 0.85、0.85 ~ 0.95 和 0.95 ~ 1 的阈值将区域划为 5 类易发区划：极低易发区划、低易发区划、中易发区划、高易发区划和极高易发区划，得到两个模型的滑坡易发性区划图，如图 6 所示。

结合研究区已知滑坡面的分布情况，并选取黄土坡滑坡、卡子湾滑坡与新滩滑坡作为参考。在滑坡易发性评价区划（图 6）结果中，RUS-SVM 耦合模型，见图 6（c）与 SMOTE-Tomek-SVM 耦合模型见图 6（d），对这三个滑坡面的预测结果吻合程度较低；对比之下，RUS-CNN 耦合模型见图 6（a）预测出的滑坡面基本吻合，SMOTE-Tomek-CNN 耦合模型见图 6（b）在此基础上，预测出的滑坡面更为吻合，表明其预测结果与实际滑坡发生面的吻合程度相较于其他耦合模型有明显提高。

4.4 试验结果对比与分析

4.4.1 特定类别精度分析

特定类别精度分析充分考虑分类区域内栅格单元个数的因素，并且可用于解决根据最易发生滑坡的区域占滑坡总面积的比例作为分析滑坡易发性评价的结果的传统方法，其所易产生两极分化的滑坡易发性评价定量分析问题[22 − 23]，该方法定义式如下：

$$p_i = \frac{A_i}{B_i} \times 100\% \tag{4}$$

式中：i=1, 2, …, n——滑坡易发性区划的分类个数；

A_i——第 i 个滑坡易发性分类区划分类中的滑坡所占栅格单元的数量；

B_i——第 i 个滑坡易发性区划分类中的栅格单元的数量；

P_i——在第 i 个滑坡易发性区划分类中的特定类别精度。

根据式（4），两模型的特定类别精度分析结果如表 4 所示。

根据表 4，经 SMOTE-Tomek 方法处理的结果（73.40%，61.17%）均表现好于传统欠采样的评价结果（64.10%，56.73%），且基于 SMOTE-Tomek 方法的 CNN 模型的评价结果是优于 SMOTE-Tomek-SVM 耦合模型的评价结果。

4.4.2 受试者工作特征（receiver operating characteristic，ROC）曲线分析

ROC 曲线是常用来验证模型性能优劣的常用指标，它可以直观地展现模型预测结果的精度和可靠性。ROC 曲线是以敏感性 TRR 为 Y 轴，以将特异性 TNR 为 X 轴，ROC 曲线越靠近左上角点时，说明分类器分类效果越好（图 7）。

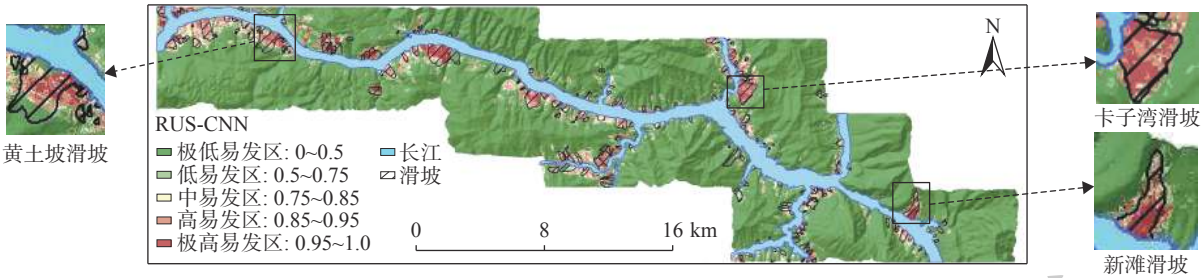
为了评价不同分类器或者分类器在不同条件下的表现时，一般是以曲线下面积（*AUC* 值）作为评价标准。

由表 5 可见，传统采样方法的 *AUC* 值均低于 SMOTE-Tomek 方法的结果，且 SMOTE-Tomek-CNN 耦合模型的 *AUC* 值为 0.965，大于 SVM 模型的 0.951。说明在 ROC 曲线分析中，基于 SMOTE-Tomek 方法与 CNN 耦合模型的滑坡易发性评价结果最优。

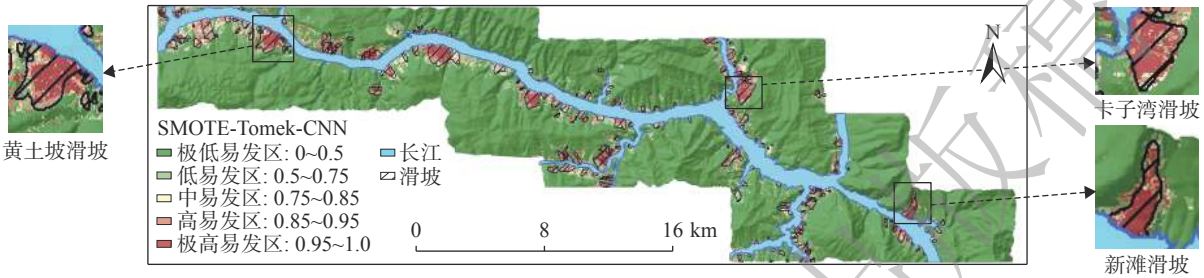
4.5 滑坡易发性评价结果分析

经过上述两种方法对滑坡易发性评价结果的分析对比，表现最好的是 SMOTE-Tomek-CNN 耦合模型，相较于传统的欠采样方法，该组合采样方法的定量分析数值有显著提升，这表明新增加的数据有与原始滑坡数据一样的预测能力，进而提高了模型的预测性能。同时，得益于 CNN 模型内部卷积核的权值共享和全连接的特点，它能充分提取隐藏的有价值的特征，并且有效防止模型产生过拟合。

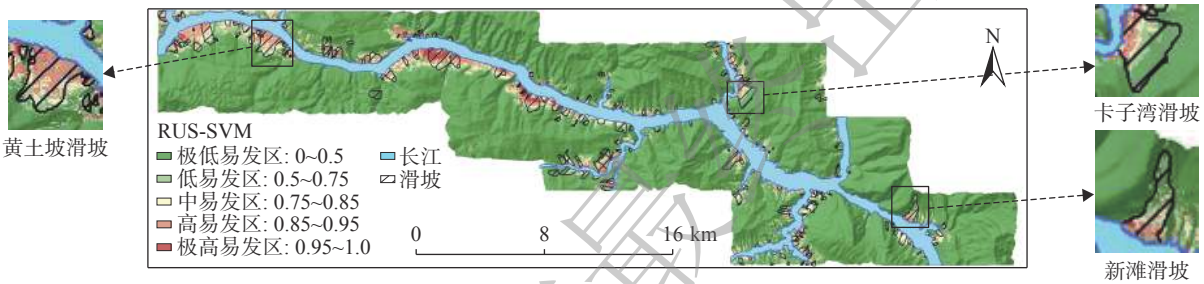
在滑坡易发性评价中，存在用来训练模型的滑坡数



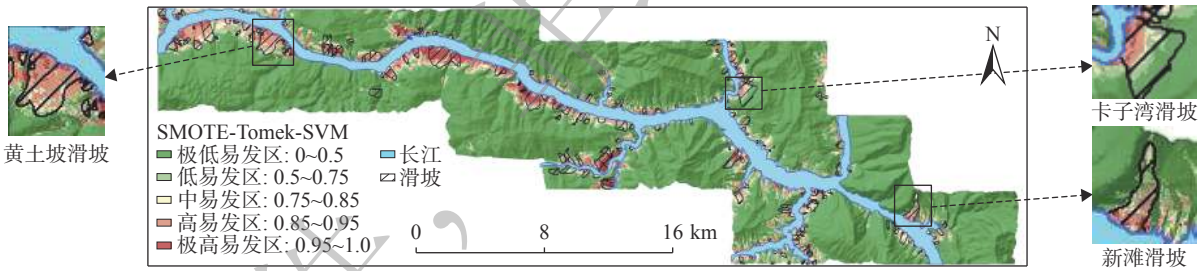
（a）基于RUS-CNN耦合模型的LSM结果



（b）基于SMOTE-Tomek-CNN耦合模型的LSM结果



（c）基于RUS-SVM耦合模型的LSM结果



（d）基于SMOTE-Tomek-SVM耦合模型的LSM结果

图 6　研究区滑坡易发性区划结果

Fig. 6　The result of landslide susceptibility assessment in the study area

据量相对较少的问题，即使传统的对非滑坡数据进行欠采样处理取得了不错的结果，但用于建模的样本量占总样本量较小，会对滑坡易发性评价预测结果的精确性和可靠性产生不利影响。SMOTE 方法生成的滑坡数据是通过线性插值得到的，滑坡与非滑坡数据在达到平衡的同时扩张了滑坡数据的数据空间，继续通过 Tomek Links 方法在经 SMOTE 方法处理过的数据集的数据空间中寻找噪声点以及边界点，增强滑坡空间与非滑坡数据空间边界的区分度，不仅使得数据数量达到平衡，也为模型提供一个更好的决策边界，提高了预测能力与分辨能力。

表 4　特定类别精度分析

Table 4　Analysis of specific category accuracy

模型	RUS-CNN	SMOTE-Tomek-CNN	RUS-SVM	SMOTE-Tomek-SVM
极低易发	1.28	0.60	0.76	0.46
低易发	15.71	16.40	9.31	9.12
中易发	27.24	29.15	23.91	24.33
高易发	41.09	45.26	38.57	44.18
极高易发	64.14	73.60	56.73	61.17

5　结论

滑坡易发性评价是对滑坡进行空间预测，同时以一

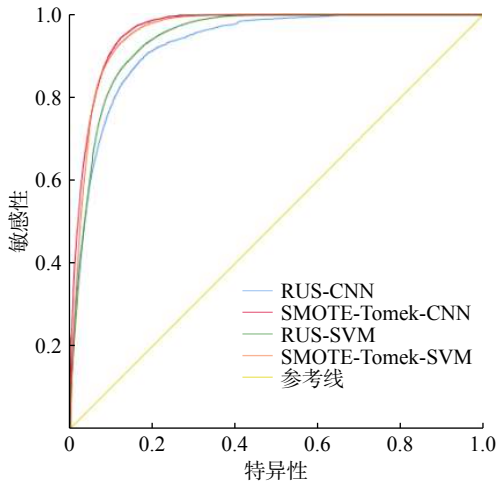


图 7 滑坡易发性评价结果的 ROC 曲线

Fig. 7 ROC curve of landslide susceptibility assessment result

表 5 曲线下面积分析

Table 5 Area under curve analysis

检验结果变量	面积	标准差①	渐进*Sig.*②	渐进95%置信区间	
				下限	上限
RUS-CNN	0.929	0.001	0.000	0.928	0.930
SMOTE-Tomek-CNN	0.965	0.000	0.000	0.964	0.965
RUS-SVM	0.942	0.000	0.000	0.941	0.943
SMOTE-Tomek-SVM	0.951	0.000	0.000	0.950	0.952

注:①在非参数假设下;②零假设:实面积=0.5。

种可视化的方式展现结果的方法。本研究除使用传统的欠采样减少非滑坡的数据量,还使用 SMOTE-Tomek 方法有效增加训练集中滑坡数据的数量,并最终使用这些新生成的滑坡样本与同等数量的非滑坡样本共同组成的平衡样本集来训练 CNN 模型与 SVM 模型,并按照一定阈值完成研究区的滑坡易区划,得到结论如下:

(1)使用 SMOTE-Tomek 方法处理过的数据集训练的模型,其评价结果均表现对处理滑坡数据不平衡有效果,通过比对 SMOTE-Tomek-CNN 与 SMOTE-Tomek-SVM 与传统欠采样方法易发性评价的结果,该方法不仅有效增加滑坡数据,还增强了滑坡数据与非滑坡在数据空间中的区分度,从而提高模型的分类与预测能力。

(2)根据特定类别精度分析与 ROC 曲线分析的结果,采用 CNN 模型,其预测结果均优于 SVM 模型。通过将一维数据转化为二维矩阵,使得 CNN 模型有效地提取滑坡空间信息,并通过共享权重来显著减少神经网络参数的数量,逐渐在因子向量中学习更复杂的特征表示,其强大的泛化能力与数据特征提取能力在未来滑坡易发性评价中有更广阔的运用空间。

(3)对于滑坡易发性区划,通过比对各已发生的滑坡面,SMOTE-Tomek 与 CNN 耦合模型预测的极高易发性区划与黄土坡滑坡、卡子湾滑坡和新滩滑坡等滑坡面吻合程度高,验证了该评价结果可靠,能为研究区滑坡预测工作提供参考。

参考文献(References):

[1] XI Chuanjie, HAN Mei, HU Xiewen, et al. Effectiveness of newmark-based sampling strategy for coseismic landslide susceptibility mapping using deep learning, support vector machine, and logistic regression[J]. Bulletin of Engineering Geology and the Environment, 2022, 81(5): 174.

[2] 于宪煜,胡友健,牛瑞卿. 基于 RS-SVM 模型的滑坡易发性评价因子选择方法研究[J]. 地理与地理信息科学, 2016, 32(3): 23 − 28. [YU Xianyu, HU Youjian, NIU Ruiqing. Research on the method to select landslide susceptibility evaluation factors based on RS-SVM model[J]. Geography and Geo-Information Science, 2016, 32(3): 23 − 28. (in Chinese with English abstract)]

[3] 贾雨霏,魏文豪,陈稳,等. 基于 SOM-I-SVM 耦合模型的滑坡易发性评价[J]. 水文地质工程地质, 2023, 50(3): 125 − 137. [JIA Yufei, WEI Wenhao, CHEN Wen, et al. Landslide susceptibility assessment based on the SOM-I-SVM model[J]. Hydrogeology & Engineering Geology, 2023, 50(3): 125 − 137. (in Chinese with English abstract)]

[4] 王雪冬,张超彪,王翠,等. 基于 Logistic 回归与随机森林的和龙市地质灾害易发性评价[J]. 吉林大学学报(地球科学版), 2022, 52(6): 1957 − 1970. [WANG Xuedong, ZHANG Chaobiao, WANG Cui, et al. Geological disaster susceptibility in Helong City based on logistic regression and random forest[J]. Journal of Jilin University (Earth Science Edition), 2022, 52(6): 1957 − 1970. (in Chinese with English abstract)]

[5] 杨得虎,朱杰勇,刘帅,等. 基于信息量、加权信息量与逻辑回归耦合模型的云南罗平县崩滑灾害易发性评价对比分析[J]. 中国地质灾害与防治学报, 2023, 34(5): 43 − 53. [YANG Dehu, ZHU Jieyong, LIU Shuai, et al. Comparative analyses of susceptibility assessment for landslide disasters based on information value, weighted information value and logistic regression coupled model in Luoping County, Yunnan Province[J]. The Chinese Journal of Geological Hazard and Control, 2023, 34(5): 43 − 53. (in Chinese with English abstract)]

[6] 杜国梁,杨志华,袁颖,等. 基于逻辑回归–信息量的川藏交通廊道滑坡易发性评价[J]. 水文地质工程地质, 2021, 48(5): 102 − 111. [DU Guoliang, YANG Zhihua, YUAN Ying, et al. Landslide susceptibility mapping in the Sichuan-Tibet traffic corridor using logistic regression-

information value method［J］. Hydrogeology & Engineering Geology, 2021, 48(5): 102 − 111.（in Chinese with English abstract）］

［7］ 林荣福，刘纪平，徐胜华，等. 随机森林赋权信息量的滑坡易发性评价方法［J］. 测绘科学，2020, 45(12): 131 − 138.［LIN Rongfu, LIU Jiping, XU Shenghua, et al. Evaluation method of landslide susceptibility based on random forest weighted information［J］. Science of Surveying and Mapping, 2020, 45(12): 131 − 138.（in Chinese with English abstract）］

［8］ 张明岳，李丽敏，温宗周. RNN 与 LSTM 方法用于滑坡位移动态预测的研究［J］. 人民珠江，2021, 42(9): 6 − 13.［ZHANG Mingyue, LI Limin, WEN Zongzhou. Research on RNN and LSTM method for dynamic prediction of landslide displacement［J］. Pearl River, 2021, 42(9): 6 − 13(in Chinese with English abstract）］

［9］ 顾华奇，陈皆红，李婷. 基于深度学习的滑坡监测与早期预警方法研究［J］. 江西科学，2019, 37(2): 209 − 213.［GU Huaqi, CHEN Jiehong, LI Ting. Research on landslide monitoring and early warning based on depth learning［J］. Jiangxi Science, 2019, 37(2): 209 − 213.（in Chinese with English abstract）］

［10］ Huijuan, ZHANG. Combining a class-weighted algorithm and machine learning models in landslide susceptibility mapping: A case study of Wanzhou section of the Three Gorges Reservoir, China［J］. Computers & Geosciences, 2022, 158: 104966.

［11］ WANG Yi. Comparison of convolutional neural networks for landslide susceptibility mapping in Yanshan County, China［J］. Science of the Total Environment, 2019, 666: 975 − 993.

［12］ XU Chong. GIS-based support vector machine modeling of earthquake-triggered landslide susceptibility in the Jianjiang River watershed, China［J］. Geomorphology, 2012, 145/146: 70 − 80.

［13］ CHAWLA N V, BOWYER K W, HALL L O, et al. SMOTE: synthetic minority over-sampling technique［J］. Journal of Artificial Intelligence Research, 2002, 16: 321 − 357.

［14］ 武雪玲，杨经宇，牛瑞卿. 一种结合 SMOTE 和卷积神经网络的滑坡易发性评价方法［J］. 武汉大学学报(信息科学版)，2020, 45(8): 1223 − 1232.［WU Xueling, YANG Jingyu, NIU Ruiqing. A landslide susceptibility assessment method using SMOTE and convolutional neural network［J］. Geomatics and Information Science of Wuhan University, 2020, 45(8): 1223 − 1232.（in Chinese with English abstract）］

［15］ 王忠洋. 基于 SMOTE-Tomek 和卷积神经网络的入侵检测模型研究［D］. 南昌：江西师范大学，2020.［WANG Zhongyang. Research on intrusion detection model based on SMOTE-tomek and convolutional neural network[D]. Nanchang: Jiangxi Normal University, 2020.（in Chinese with English abstract）］

［16］ 于宪煜. 基于多源数据和多尺度分析的滑坡易发性评价方法研究［D］. 武汉：中国地质大学，2016.［YU Xianyu. Study on landslide susceptibility evaluation method based on multi-source data and multi-scale analysis[D]. Wuhan: China University of Geosciences, 2016.（in Chinese with English abstract）］

［17］ YU Xianyu, ZHANG Kaixiang, SONG Yingxu, et al. Study on landslide susceptibility mapping based on rock–soil characteristic factors［J］. Scientific Reports, 2021, 11: 15476.

［18］ LI Wenjuan, FANG Zhice, WANG Yi. Stacking ensemble of deep learning methods for landslide susceptibility mapping in the Three Gorges Reservoir area, China［J］. Stochastic Environmental Research and Risk Assessment, 2022, 36(8): 2207 − 2228.

［19］ FANG Zhice, WANG Yi, PENG Ling, et al. A comparative study of heterogeneous ensemble-learning techniques for landslide susceptibility mapping［J］. International Journal of Geographical Information Science, 2021, 35(2): 321 − 347.

［20］ DAHL G E, SAINATH T N, HINTON G E. Improving deep neural networks for LVCSR using rectified linear units and dropout［C］//2013 IEEE International Conference on Acoustics, Speech and Signal Processing. May 26-31, 2013, Vancouver, BC, Canada. IEEE, 2013: 8609 − 8613.

［21］ KINGMA D P, BA J. Adam: A method for stochastic optimization[EB/OL]. 2014: arXiv: 1412.6980.

［22］ YU Xianyu, GAO Huachen. A landslide susceptibility map based on spatial scale segmentation: A case study at Zigui-Badong in the Three Gorges Reservoir Area, China［J］. PLoS One, 2020, 15(3): e0229818.

［23］ SÜZEN M L, DOYURAN V. A comparison of the GIS based landslide susceptibility assessment methods: Multivariate versus bivariate［J］. Environmental Geology, 2004, 45(5): 665 − 679.